# MATH 462 LECTURE NOTES
# MULTI CLASSIFICATION

ADAM M. OBERMAN

## 1. Setup for multi class, score based

Let $K \geq 2$ be the number of classes.
Dataset

$$S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \tag{S}$$

The roadmap for $K$-classification, and this section, is as follows:

- But now we have $y \in \mathcal{Y}_K = \{1, 2, \ldots, K\}$, corresponding to $K$ classes
- We now have a vector of scores, $s(x) \in \mathbb{R}^K$, one score for each class.
- The important qualities of a loss are still important: it should be a convex, differentiable upper bound for the error (the zero-one loss).

Our classification map $c : \mathbb{R}^K \to \mathcal{Y}_K$ is defined as

$$c(s) = \arg\max_i s_i$$

As before, on a dataset $(S)$, given a score based loss, $\ell$ and function $h : \mathcal{X} \to \mathbb{R}$, the average loss of $h$ on $S$ is given by

$$L(h, S) = \frac{1}{m} \sum_{i=1}^{m} \ell(h(x_i), y_i)$$

**Definition 1.1** (Error)**.** The error, or zero-one loss, is given by $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$,

$$\ell_{0,1}(y_1, y_2) = \begin{cases} 0 & y_1 = y_2 \\ 1 & \text{otherwise} \end{cases}$$

Given a function $c : \mathcal{X} \to \mathcal{Y}$ and the dataset $(S)$, the error of the model on the dataset is given by

$$L_{0-1}(c, S) = \frac{1}{m} \sum_{i=1}^{m} \ell_{0-1}(c(x_i), y_i)$$

## 2. Multi class Losses

**Warning: confusing notation for losses**

Concise but confusing notation. $s = (s_1, \ldots, s_K)$ is a vector. Usually we use $i, j, k$ for the components. When we write $s_y$ this is the $y$th component of $s$.
E.g. $s = (1, 3, 4, 7, 8), y = 2$, so $s_y = s_2 = 3$

*Date*: October 10, 2023.

2.1. **Log-loss.** Next consider the $K$ class case with the log loss.

Given $s \in \mathbb{R}^K$, define the softmax function `https://en.wikipedia.org/wiki/Softmax_function`, $\sigma : \mathbb{R}^K \to [0,1]^K$, by

$$\sigma(s)_j = \frac{e^{s_j}}{\sum_{i=1}^K e^{s_i}}$$

Note this maps a vector to a probability vector: the sum of the components is one. $\sum_i \sigma(s)_i = 1, \quad \forall s \in R^K$.

Then define the loss

$$\ell_{\log -K}(s,y) = -\log(\sigma(s)_y) \tag{1}$$

Here $\sigma(s)$ plays the role of the probability of each class, and we have the multiclass version of the log loss.

*Example* 2.1. Notice in (1), $y$ has the role of the index. So for example, if $K = 3$, and $\sigma(s) = (.2, .7, .1)$, then $\ell_{\log -K}(s, 2) = -\log(.7)$ and $\ell_{\log -K}(s, 3) = -\log(.1) = \log(10)$

This loss is convex, but it takes a bit of work to show it. First, we can rewrite the loss as follows, For $s \in \mathbb{R}^K, y \in \mathcal{Y}_K$, define the shifted vector

$$m(s,y)_i = s_i - s_y$$

Then

$$\ell_{\log -K}(s,y) = \mathrm{LSE}(m(s,y)) = \log(\exp(m_1(s,y)) + \ldots \exp(m_K(s,y)))$$

*Example* 2.2. Next, consider $s = (0.5, 4.5, 0.5)$, and $y = 1$. For this score we have a margin of $m(s,y) = (0,4,0)$. Here, the loss is $\log(2e + e^4) > \log(3)$, so w are incorrect (and sure!)

When $y = 2$, we have $m(s,2) = (-4, 0, -4)$. Here the loss is $\log(1 + 2\exp(-4))$ which is small.

Fix $y = 1$. For our first score, take $s = (2/3, 2/3, 2/3)$. This has associated margin $m(s,y) = (0,0,0)$. For this the loss is $\log 3$. It's not clear if the classification is correct. Perturbing the scores by a small amount leads to a similar loss value.

The convexity of the loss follows from general properties of convex functions (since convex in $m$ and $m$ linear in $s$ means convex in $s$). The convexity of log-sum-exp is shown in page 72 of Boyd, [BV04].

What about upper bound for the zero one loss?

One can check this is an upper bound for the zero-one loss in the sense that

$$\ell_{0-1}(c(s), y) \leq \frac{1}{\log 2} \ell_{log-K}(s,y), \quad \forall s, y$$

However, there is one property we lose: a one-to-one correspondence between the loss value and the classification. Namely we have

$$\begin{aligned} c(s) = y &\implies \ell_{log-K}(s,y) \leq \log K \\ c(s) \neq y &\implies \ell_{log-K}(s,y) \geq \log 2. \end{aligned}$$

Or in words, when $\ell \in [\log 2, \log K]$ the classification may be correct, or incorrect.

2.2. **Margin loss.** Now, we generalize the margin loss from the binary case. We wish to do so in a way that preserves the desirable properties of the margin loss: upper bounding the zero-one loss, convexity, and the interpretability of the loss value. We also point out that the shortcomings of the margin loss should still be the same when working with $k$-classes: non-differentiability at zero.

**Definition 2.3** (Margin Loss for $K$ classes)**.** For $s \in \mathbb{R}^K, y \in \mathcal{Y}_K$, define the margin

$$\delta(s, y) = s_y - \max_{j \neq y} s_j$$

and the loss

$$\ell_{margin-K}(s, y) = \max(0, 1 - \delta(s, y)).$$

*Example* 2.4. For example, $\delta((5.2, 3, .5), 1) = 2.2$ and $\delta((5.2, 3, .5), 3) = .5 - 5.2 = -4.7$. and $\ell_{margin-K}(s, y) = 0$ in the first case, and $5.7$ in the second case.

*Example* 2.5. Note that when the two (or more) of largest scores are the same, e.g. $s = (5, 6, 6)$ then $m(s) = 0$ and the $\ell(s, 2) = \ell(s, 3) = 1$. So the loss is one on the classification boundary.

Helpfully, we have the same one-to-one correspondence between loss values and classification. That is, we still have that a loss less than one implies we have correctly classified, and that a loss greater than one means we have incorrectly classified. Precisely:

$$\ell_{margin-K}(s(x), y) \leq 1 \iff c(s(x)) = y$$

2.3. **Loss design.** We want a loss that encourages more confident (correct) classifications. This leads to the following loss design principle.

The approach to (supervised) binary classification we take is score based, differentiable loss. The main advantage of this approach is that a differentiable loss is amenable to optimization

**Definition 2.6** (Loss design principle)**.** We still want the loss to be an upper bound for the error

$$\ell(s, y) \geq \ell_{0-1}(c(s), y), \quad \forall s, y \tag{2}$$

We say the loss is:

(1) Balanced, if permuting labels and scores doesn't change the loss.
(2) An upper bound for the error, if (2) holds.
(3) Normalized if $\ell(s, y) = 1$ for some $s$ on the classification boundary.
(4) monotone if $\ell(s, y)$ is increasing in $s_y$ and descreasing in $s_j$ for $j \neq y$.
(5) convex if $\ell(s, y)$ as a function of $s$.

## References

[BV04] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.